Education Endowment Foundation

Guidance for evaluators and funders on using GCSE performance as an outcome measure

EEF Research Paper

July 2021

Authors:

Ben Smith, Stephen Morris and Harry Armitage

AlphaPlus

Manchester Metropolitan University

Summary of guidance

- Evaluators should always consider the 'minimally important effect size', i.e. what difference in mean performance would have to emerge between treatment and control groups for the intervention to be considered worthwhile?
    - This can be rephrased in more pragmatic terms as "what proportion of students would the intervention need to improve the grade of for the intervention to be considered worthwhile?"
- Evaluators should always consider the plausibility of the difference implied by the minimally important effect size/selected MDES. Is it reasonable that the intervention group will gain X grades/marks on average as a result of the intervention?
    - How proximal or distal an intervention's embedded mechanism of change is to the outcome measure will likely have a considerable impact on the difference's plausibility. Considering the specific question types and content the outcome measure's assessments cover may be useful in establishing this.
    - What ability level of candidates the intervention targets should also be considered, as this impacts on the plausibility of this difference too. The impact of capping effects (especially in tiered subjects) and mark scheme hurdles should be considered.
- Broadly, evaluators should always seek to gather mark data to use as the outcome measure, rather than relying on grades. This is because marks are finer-grained, and ultimately interventions must be difference-making on the level of marks; grades are merely a summary of marks. If however grades are the only option (or if marks are much more challenging to acquire and evaluators wish to determine the likely impact of using grades instead of marks) then the following can be considered:
    - Evaluators can consider the 'grade boundary span' of the qualification used as their outcome measure. Broadly, the more marks within each grade, the more sensitivity/power is lost if grades are used as the outcome measure.
    - Evaluators can attempt to quantify the loss in sensitivity by using freely available data and simulation techniques to derive the mark and grade distribution of the qualification used as their outcome measure. Ultimately, the 'ratio of [the SD of the grade distribution] to [the SD of the mark distribution, converted into grades]' is what dictates whether sensitivity/power is lost when grades are the outcome measure.

Guidance for evaluators & funders on using GCSE performance as an outcome measure

This guidance is intended for the planning stage of trials – in particular, trials for which the primary outcome measure being considered is attainment at GCSE (and in particular, GCSE English Language and Mathematics). However, it is generalisable to any other graded qualification or assessment used as an outcome measure, i.e. A-levels.

It is also important to note that some of the points in this guidance may not strictly be under the control of the evaluator of an intervention – hence titling it for "evaluators and funders". Some decisions may be made before an evaluator begins their work, when selecting which interventions go to trial (though arguably the evaluator of a pilot should be making recommendations in the pilot report on these points).

Approach to determining a sample size

The guidance applies largely to the determination of the sample size needed for a given trial. Typically, there are two approaches to this: 1) select an MDES value which theory and prior expectations suggest the intervention should be capable of producing, which in turn allows the selection of a sample size in accordance with this and 80% power; or 2) determine a minimally important effect size necessary for the intervention to be considered 'worthwhile', set it as the MDES and determine a sample size as above.

We would contend that it is always important to consider the minimally important effect size, even if largely following approach 1. This is because it forces the consideration of the question "what per cent of students need to gain a grade for the intervention to be worthwhile?" – which is ultimately what defines what an 'important' difference is in the context of any given intervention using GCSE performance as an outcome measure. For example, an improvement of a full grade (on average) for all participating students would clearly render a trial valuable unless it were prohibitively expensive, but only 1/30 students being expected to gain a grade may not.

It is likely to prove challenging to precisely define the values of X in the statement above at this stage. Quite possibly, not until working through to the end of this guidance will the reader be equipped to precisely define these values. However, the above question is still extremely important to be thinking about up front. This is because the concept of the 'minimally important' difference or effect size underpins all subsequent points, and serves to initiate a consideration of the nature of the outcome measure and what the numeric values it takes actually "mean" in real terms.

What are the 'plausible gains' the intervention could cause?

Having considered the minimally important difference, it is also important to consider whether the difference in GCSE performance between intervention and control groups the minimally important difference implies is plausible – that is, whether the intervention itself could reasonably lead to an improvement in GCSE grades of this magnitude. Here the key consideration is how the intervention aims to act upon performance via its associated theory of change (and any other prior evidence). This is vital as ultimately, if the implied difference is not plausible, one must question whether the intervention itself is credible.

Often the main consideration will be how proximal the mechanisms of change embedded in the intervention are to specific elements of examinations. For example, a tutoring intervention does not specifically target any particular element of an examination, whilst an intervention aiming to improve numerical reasoning would target specific types of question in Mathematics exams. There may be a combination of these effects; a numerical reasoning intervention may impact considerably on reasoning items but have a lesser effect on other types of items.

If the mechanisms of change embedded in an intervention act proximally on specific elements of the examination(s) used as outcome measures, then the freely available sample assessment materials awarding organisations host online can be used to determine how many items/marks assess the targeted skills. If an intervention solely targets these skills, then this exercise serves to indicate the upper limit on plausible marks gains the intervention can lead to.

A further consideration is that, if an intervention is particularly specific about the marks it 'targets', whether it makes sense to gather data on students' marks on those specific items. For example, an intervention targeted at improving spelling and grammar may only affect the handful of SPAG marks available on GCSE English Language papers. Whilst grades are ultimately the unit of performance that 'matters' to stakeholders, for very targeted interventions it might be extremely challenging to detect an effect in grade units. As such, knowing marks on targeted items could be valuable to derive a secondary outcome variable which more directly measures the targeted skills/competencies.

If an intervention's embedded mechanisms of change are instead more distal to the examination, then the key consideration is how the intervention can act to improve performance on the exam. This should be covered via the intervention's theory of change. Generally, we would suggest that such 'distal' interventions are likely to result in relatively small mark gains. Evaluators should review reports from prior trials (of similar interventions, where possible) to see what level of mark gains on examinations have manifested historically from other 'distal' interventions.

Even if the minimally important difference in grades is deemed plausible, there is one further complication that must be considered. As alluded to above, whilst GCSE grades are the 'meaningful unit' for stakeholders, grades are entirely determined by underling marks. Grade gains manifest because an intervention improves the average marks the treatment group achieves relative to average marks gained among a control group. As a result, it is therefore always essential to consider whether the minimally important difference is plausible if considered in terms of marks, as opposed to grades – even if grades are ultimately used as the outcome measure.

What ability level of candidates does the intervention target?

One important factor which may impact on the plausibility of the minimally important difference is students' ability level, because most interventions are unlikely to act on all students equally (indeed this is an important factor in closing the attainment gap; we implicitly hope that interventions help students from deprived backgrounds more than those from affluent ones). For example, an intervention aiming to improve students' analysis skills may not be useful for lower ability students stuck at the knowledge 'hurdles' within GCSE levels of response mark schemes (if students do not remember the relevant content, they cannot progress to demonstrate higher order skills such as analysis).

As such, it is also important to consider what ability level of students the intervention is likely to act on, and indeed this is an element mentioned in the theory of change guidance EEF provides (this guidance can be found **here**). The answer is unlikely to purely be high or low ability students, but an intervention may act on different parts of the ability spectrum to greater or lesser extents.

A key factor to be aware of is that interventions aimed at high ability students necessarily have less capacity to improve marks; such students are already scoring highly on the exam overall and on many items. The logical extreme is that if a student would already achieve maximum marks on the items an intervention acts on without it, then its impacts could not be detected. Lower ability students have much greater 'room to improve'.

However, this is complicated for tiered subjects. Here, there is an additional 'cap' on performance at the top of the Foundation tier, for students in the middle of the ability range. Unless the intervention changes entry pattern behaviour, shifting top-of-Foundation students to the Higher paper, an additional capping effect may manifest for students of middling ability.

As stated above, ultimately this acts upon the plausible mark gains we can expect students to make. For example, an intervention targeted at improving higher-order analysis skills may be expected to improve outcomes more for higher ability students than lower ability students.

Are marks or grades a more appropriate outcome measure?

As alluded to above, grades are a more coarse measure of performance than marks, and not all gains in marks will translate into grade gain. As such, it is also important to consider whether marks are a more sensitive outcome measure than grades; whether it is 'easier' to detect an effect a given number of marks in size when marks or grades are the outcome measure. There are two key factors here:

- The tariff available for the subject. More marks means marks are increasingly finer grained relative to grades, and are thus more likely to be more sensitive as an outcome measure.
- The number of grades awarded (which is higher for double award GCSEs and lower for tiered subjects). More grades awarded means more grade boundaries, which have to fit into the same tariff and thus impact how coarse grades are as a measure relative to marks.

These two factors combine to influence how far apart the grade boundaries are (though the distribution of marks candidates achieve will also affect this). Grade boundary span (measured in marks) is what ultimately affects how fine-grained marks are relative to grades, and thus how sensitive marks are relative to grades.

However, all grade boundaries within a subject are rarely precisely evenly spaced. Whilst it is possible to take an 'average boundary width', this overlooks the fact that some grades are achieved by students more often than others. To provide an extreme example using GCSE, perhaps grade 9 is only three marks wide, but grade 5 is twenty marks wide. Vastly fewer students achieve a grade 9, whilst grade 5 is one of the most common grades awarded in most subjects – so grade 5's width is of much more 'impact' than grade 9's. As such, when determining the impact grade boundary width has on whether marks are a more sensitive outcome measure, it's important to place slightly more weight on the 'more used' grades' widths.

Ultimately, our advice would be for all evaluators using GCSE performance as an outcome measure to gather examination marks. This is because the intervention must always be difference-making on the underlying mark distribution, even though grades are the more 'important' metric to most stakeholders. As outlined above, it is also more straightforward to assess the effects of different potential MDES and sample sizes on marks than grades because the latter is a summary metric of the former, and the conversion to grades is not straightforwardly linear.

If grades are the only option for outcome measurement of GCSE performance for whatever reason, then evaluators will still need to consider the plausibility of the minimally important difference in marks as well as in grades as discussed above. It also becomes important to understand the ramifications for the sample size required of the use of grades as the outcome measure; the following section details how this can be done more quantitatively than the consideration of grade boundary spans outlined above.

Quantitatively assessing whether marks or grades are more sensitive as an outcome measure

The prior section highlights some broad rules of thumb on establishing whether marks may be more sensitive as an outcome measure. However, this can also be examined more formally. The question of whether it is easier to detect an effect of given magnitude with marks as the outcome variable is ultimately down to the 'ratio of [the SD of the grade distribution] to [the SD of the mark distribution, converted into grades]'.

By way of explanation, consider a hypothetical examination where the SD of the grade distribution is 0.5 grades, and the SD of the mark distribution is 10 marks. Because effect sizes are standardised, a

trial powered to an MDES of 1.0 would be able to detect an effect either 0.5 grades in magnitude, or 10 marks in magnitude 80 per cent of the time (with 0.8 power). But are 10 marks equivalent to 0.5 grades? As outlined above, this is dictated by the grade boundary widths. We present two hypotheticals to exemplify the issue (though these gloss over the distribution of students across grades of varying widths for simplicity).

- If the average grade width is 5 marks in size, then the two SDs are exactly equivalent; one SD in marks, when converted into grades, is exactly equivalent to the SD of the grade distribution. This indicates that there is no loss in sensitivity entailed in using grades as the outcome variable.
- If the average grade width is 10 marks in size, then the two SDs are not 'equivalent'. One SD of marks, if converted into grades, would be double the SD of grades. This indicates that there is a substantial loss in sensitivity entailed in using grades as the outcome variable.

In the former case, because grades are just as sensitive as marks, it is reasonable to use grades as the outcome variable with little further consideration of the issue. However, in the latter case, there would be a substantial effect of using grades rather than marks as an outcome variable. In effect, a substantially larger sample size would be required in order to detect the same effect if grades are the outcome variable, with commensurate cost implications. It may be useful to weigh these up against the costs associated with gathering mark information from a smaller sample.

The information required to deduce any GCSE subject's grade distribution (and thus the SD of this distribution) is freely available online; JCQ results statistics report the number of students achieving each grade for every qualification. Mark distributions are however not freely available. That said, **Smith, Boyle & Morris (2020)** used freely available tariff and grade boundary information from awarding organisations' websites (in conjunction with grade distribution data) to simulate the mark distribution for GCSE Science; similar methodology can be applied to other subjects with relative ease to approximate the SD of marks.

Smith, Morris & Armitage (2021) investigated the above specifically for GCSE English Language and GCSE Mathematics, the two most common GCSEs used as outcome measures in evaluations. In short, their findings were that in GCSE English Language grades were almost exactly as sensitive as marks, whilst in GCSE Mathematics there was a significant loss in sensitivity associated with using grades as an outcome variable, largely due to Maths' higher tariffs and tiered structure meaning there are fewer grades available for each separate tier.

Note that having carried out this exercise, the reader will be armed with the information needed to define X in our prior statement of the 'minimally important difference': "X per cent of students need to gain a grade". The average treatment effect associated with the MDES the trial is powered to can be deduced once the SD of the grade distribution is known – from freely available information. For example, if the SD of the grade distribution is 0.5, and the trial is powered to detect an effect 0.25 SDs in size, then our minimally important difference is that "1/8 students need to gain a grade".

In effect, the "minimally important difference" is to some degree a restatement of the MDES the trial is powered to detect into meaningful units. However, it is not entirely defined by the MDES – it may be that having done the above calculations, the emerging minimally important difference feels "wrong" for some other reason. This is entirely legitimate – there are myriad factors that determine what an "important difference" is. We can use the SD and MDES to derive the "minimum detectable difference", but there is always a degree of subjective or qualitative input in deciding whether this is an "important difference" or not.

This document is available for download at **www.educationendowmentfoundation.org.uk**

Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk